Inflexible social inference in individuals with subclinical persecutory delusional tendencies

Katharina V. Wellstein[1] ¶*, Andreea Oliviana Diaconescu[1,2,3] ¶, Martin Bischof[4], Annia Rüesch[4,5], Gina Paolini[1,6], Eduardo A. Aponte[1], Johannes Ullrich[5], & Klaas Enno Stephan[1,7,8]

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Switzerland
[2] Department of Psychiatry (UPK), University of Basel, Switzerland
[3] Krembil Centre for Neuroinformatics (CAMH), University of Toronto, Canada
[4] Department of Psychiatry, University Hospital of Psychiatry (PUK), University of Zurich, Switzerland
[5] Department of Psychology, University of Zurich, Switzerland
[6] Klinik für Psychiatrie und Psychotherapie, Clienia Schlössli AG
[7] Wellcome Centre for Human Neuroimaging, University College London, UK
[8] Max Planck Institute for Metabolism Research, Cologne, Germany

[¶] These authors contributed equally to this work

Corresponding Author

Katharina V. Wellstein
Translational Neuromodeling Unit
University of Zurich & ETH Zurich
Wilfriedstrasse 6
CH – 8032 Zurich

0041 44 634 91 11
wellstein@biomed.ee.ethz.ch

Keywords
psychosis, delusion, dimensional psychiatry, social cognition, inference, persecutory ideation, social learning

1. Abstract

It has been suspected that abnormalities in social inference (e.g., learning others' intentions) play a key role in the formation of persecutory delusions (PD). In this study, we examined the association between subclinical PD and social inference, testing the prediction that proneness to PD is related to altered social inference and beliefs about others' intentions. We included 151 participants scoring on opposite ends of Freeman's Paranoia Checklist (PCL). The participants performed a probabilistic advice-taking task with a dynamically changing social context (volatility) under one of two experimental frames. These frames differentially emphasized possible reasons behind unhelpful advice: (i) the adviser's possible intentions (dispositional frame) or (ii) the rules of the game (situational frame). Our design was thus 2x2 factorial (high vs. low delusional tendencies, dispositional vs. situational frame). We found significant group-by-frame interactions, indicating that in the situational frame high PCL scorers took advice less into account than low scorers. Additionally, high PCL scorers believed more frequently that incorrect advice was delivered intentionally and that such misleading behaviour was directed towards them personally. Overall, our results suggest that social inference in individuals with subclinical PD tendencies is shaped by negative prior beliefs about the intentions of others and is thus less sensitive to the attributional framing of adviser-related information. These findings may help future attempts of identifying individuals at risk for developing psychosis and understanding persecutory delusions in psychosis.

2. Introduction

Delusions represent a hallmark of psychosis and are conceptualized as false beliefs based on incorrect inference about the external world that persist in the face of disconfirmatory evidence (DSM-IV 2000, p. 765 and DSM-5 2013, p. 819). The most prominent delusional beliefs pertain to the social world, specifically that other indviduals' intentions are of persecutory nature (Bell & Halligan, 2013). Persecutory delusions (PD) shape the experience of 70% of first episode psychosis patients and 50% of patients from the schizophrenia spectrum(Freeman, 2007; Freeman & Garety, 2014; Sartorius et al., 1986), and are related to reduced psychological well-being (Freeman et al., 2014) and higher risks of violence (Keers, Ullrich, DeStavola, & Coid, 2014).

The understanding of delusions as abnormal beliefs and their immunity to disconfirmatory evidence led to influential concepts that suggested abnormalities of Bayesian inference (i.e. inference based on integrating observations with prior beliefs) as the cause of delusion formation (Coltheart, Menzies, & Sutton, 2010; Hemsley & Garety, 1986). The notion that delusional ideation may be associated with abnormal inference has previously been related to the Jumping to Conclusions (JTC) bias in delusions (e.g., (Garety, Hemsley, & Wessley, 1991; Peters & Garety, 2006; So et al., 2012; Speechley, Whitman, & Woodward, 2010); but see (Ermakova, Gileadi, Knolle, Diaz, & Anderson, 2017; Moutoussis, Bentall, El-Deredy, & Dayan, 2011) for alternative interpretations).

A more recent Bayesian account of delusions refers to the interplay between prior beliefs and prediction error (PE) signals (Corlett, Honey, & Fletcher, 2016; Corlett, Taylor, Wang, Fletcher, & Krystal, 2010; Fletcher & Frith, 2009; Sterzer et al., 2018). This derives from one prominent Bayesian perspective on perception – predictive coding (Friston, 2005; Rao & Ballard, 1999) – which proposes that the brain infers the causes of its sensations using a hierarchical model of the external world. This model is assumed to represent beliefs that provide top-down predictions about sensory inputs, which are then adjusted by PEs at each level of the hierarchy. According to hierarchical Bayesian schemes, delusion formation might reflect a compensatory response to deficiencies of hierarchical inference (Adams, Stephan, Brown, Frith, & Friston, 2013; Corlett et al., 2016; Fletcher & Frith, 2009). Specifically, delusions might result from efforts to "explain-away" abnormally precise low-level PEs leading to adaptation of beliefs at higher levels in the processing hierarchy (Adams et al., 2013; Schmack et al., 2013). If these PEs are "chaotic" and result from usually unremarkable events (cf. "aberrant salience"; (Heinz, 2002; Kapur, 2003; Shaner, 1999)), adopting general and overly precise higher-level beliefs, may be a way of making sense of these events.

Adopting precise higher-level beliefs is crucial in social contexts, since human intentions are typically concealed (Biedermann, Frajo-Apor, & Hofer, 2012). Previous studies linking Theory of Mind (ToM) and PD found that patients with PD have difficulty taking contextual factors into account when thinking about others' intentions and exhibit an enhanced externalising bias (i.e., a tendency to blame others rather than the situation for negative events, see (Craig, Hatton, Craig, & Bentall, 2004; Langdon,

Corner, McLaren, Ward, & Coltheart, 2006; Lincoln, Mehl, Exner, Lindenmeyer, & Rief, 2010); but also see (Martin & Penn, 2002; McKay, Langdon, & Coltheart, 2005)). This tendency to attributing harmful intent to others has also been associated with subclinical persecutory ideation (Raihani & Bell, 2017, 2018).

In this study, we investigated social inference in individuals with subclinical PD tendencies. Based on the notion of rigid high-level beliefs playing a crucial role in PD and assuming that PD is a dimensional construct, we generally expected that (1) individuals with subclinical tendencies towards PD should behave less adaptively in dynamic social contexts and (2) should be less sensitive to experimental manipulations of attributional biases (experimentally-induced priors).

Adopting a dimensional perspective on PD (Van Os et al., 1999), we invited participants from the general population scoring either high or low on the Paranoia Checklist (PCL; (Freeman et al., 2005; Lincoln, Ziegler, Lüllmann, Müller, & Rief, 2010)). To investigate social inference, we employed an iterative probabilistic advice-taking paradigm. We probed the influence of attributional priors on social inference by introducing two experimental frames with minor differences in how the cause of incorrect advice was framed: First, a dispositional frame served to direct participants' attention to the adviser's character – namely, that the adviser acted intentionally in order to achieve his/her own (unknown) goals. Second, a situational frame directed participants' attention to the contextual aspects of the task – namely, that the adviser's behaviour was not only influenced by his/her intentions but also by the incompleteness of the information available to him/her. Individuals with relatively agnostic beliefs about the adviser were expected to learn and adhere to advice differently depending on how the task was framed. By contrast, individuals with proneness to PD were expected to be governed more by their own high-level beliefs than task-induced attributional priors. Thus, we predicted group-by-frame interactions in participants' decisions to adhere to advice during our social inference paradigm (Hypothesis I). Furthermore, we expected that high PCL scorers vs. low PCL scorers would attribute incorrect advice *less* to the adviser having incomplete information (Hypothesis II), and rather attribute negative events (e.g. bad performance on the task) to external-personal causes (Hypothesis III). Specifically, we hypothesized that high PCL scorers would attribute incorrect advice *more* to the adviser as a person than to themselves or the social context (Hypothesis IV). Additionally, we predicted that high PCL scorers would expect misleading advice (Hypothesis V) and believe that the adviser's giving incorrect advice was directed toward them personally (Hypothesis VI). For an overview of the hypotheses, their operationalization, and the results see Supplementary Figure 1 in the Supplementary Material.

These hypotheses were defined in an analysis plan prior to data analysis (https://gitlab.ethz.ch/sibak/sibak-analysis-plan). For a summary of all hypotheses and results, please see Supplementary Figure 1. Notably, the hypotheses addressed in this paper refer exclusively to behavioral readouts from the task or to self-report measures. An independent analysis of the behavioural

data that adresses additional hypotheses using a computational model of learning and inference is presented in a separate paper (Diaconescu, Wellstein, Mathys, & Stephan, 2019).

3. Materials and Methods

3.1 Pre-Screening

Participants were recruited from the general population via online platforms for students and locals and via print adverts at stores. *N*=1,145 individuals were pre-screened online with the Paranoia Checklist (PCL; (Freeman et al., 2005; Lincoln, Ziegler, et al., 2010), intermixed with distractor items from the NEO-FFI (McCrae & Costa, 2004)), allowing us to assign individuals to groups characterised by high ("high PD") or low tendencies towards PD ("low PD"). The PCL is a self-report questionnaire consisting of 18 items representing statements linked to paranoid ideation. Subscales include frequency of the thoughts occurring, conviction, and distress. Group assignment was based on the frequency subscale. See Figure 1 for the distribution of PCL frequency scores in our pre-screened sample. The inclusion criteria for participating in the online pre-screening were as follows: (i) age 18 or older, (ii) fluent German comprehension, and (iii) absence of current treatment. The groups were defined in reference to the mean and standard deviation of the PCL scores obtained in healthy volunteers by Freeman et al. (Freeman et al., 2005). We used these reference values in order to enable continuous inclusion of participants during ongoing prescreening. Participants scoring 0.5*sd* above this mean were assigned to the high PD group and those scoring 0.5*sd* below were assigned to the low PD group.

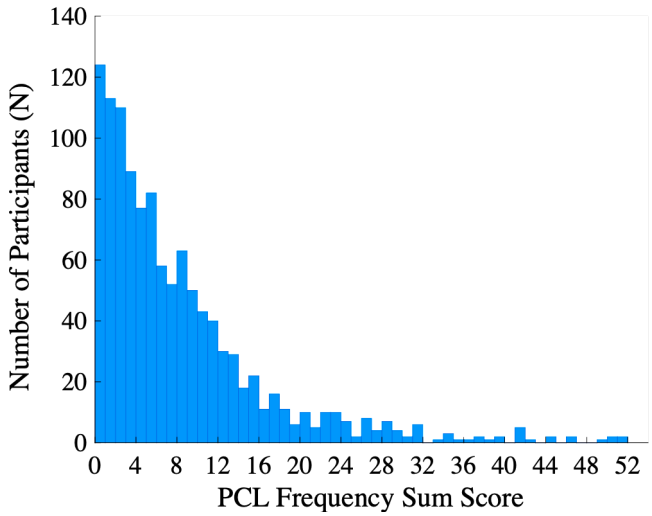

Figure 1 | **Histogram of PCL frequency scores in pre-screening sample**

The PCL frequency subscale assesses frequency of paranoid thoughts (18 items), with 0="*less than once every month*" and 4="*at least once a day*"; the highest possible score is 72.

N=1'145 individuals filled in the questionnaire during pre-screening.

This sample:
mean=7.73, *median*= 5.0, *sd*=8.39, *range*=0-52.

Measures reported by Freeman et al. (2005):
mean=11.9, *median*= 9.0, *sd*=10.5, *range*=0-64.

In order to reduce the possibility of group assignment being based on a transient expression of persecutory ideation, participants assigned to one of the two groups were invited to fill in the online questionnaire again (screening), four weeks after participating in the pre-screening, which was the case

for N=344. Only individuals whose score remained outside the 0.5 *sd* intervals described above when completing the PCL for the second time were invited to the experiment, where they completed the PCL for a third time. The latter served to exclude any systematic differences in PCL scores obtained online versus on site.

3.2 Sample

We used a 2x2 factorial between-subject design with two participant groups assigned pseudo-randomly to two experimental conditions. Based on a power analysis (using g-Power (Faul, Erdfelder, Lang, & Buchner, 2007)), a sample size of N=146 was computed for a minimum of 80% power at α=0.05 under a moderate effect size (Cohen's f^2=0.25), required to run a two- and three-way ANOVA. Concerning the effect size, while there are no previous results that exactly relate to our questions, we sought guidance by a previous study (Diaconescu et al., 2014) which used the same type of task and reported a moderate effect size for a related question.

Assuming a drop-out rate of 10% (based on studies using the same task by Diaconescu et al. (Diaconescu et al., 2014, 2017)), we invited 162 participants to the experiment, and matched low PD and high PD participants regarding age, education, and proportion of male vs. female. Eleven participants were excluded from analyses based on previously defined exclusion criteria (specified in an analysis plan, time-stamped before completion of data acquisition and analyses (https://gitlab.ethz.ch/sibak/sibak-analysis-plan).

Only participants with an average score over all three questionnaires (pre-screening, screening, and experiment day) that was outside the ±0.5 *sd* intervals were included in the analyses, which was the case for 151 of the 162 individuals.

3.3 Experiment

3.3.1 Experimental procedure

After providing informed consent, participants received written standardised task instructions. To ensure that they understood the task, they were asked to explain it in their own words and performed a practice round (8 trials) in which they were truthfully informed that advice validity was fixed to chance.

After completing the task, participants filled out a task-specific debriefing questionnaire and were administered a cognitive screening – the symbols-test (coding) and the letters-numbers test (working memory) of the Brief Neurocognitive Assessment (BNA; (Fervaha et al., 2014)) to control for the potential influence of cognitive deficits on social inference. (Ventura, Wood, & Hellemann, 2013) It was administered after the task to avoid possible influences of cognitive load on social inference (Gilbert & Osborne, 1989). After filling out the PCL again (Freeman et al., 2005; Lincoln, Ziegler, et al., 2010), participants were reimbursed and debriefed about the study before they left.

3.3.2 Task

The task used in this study is a modified version of the advice-taking task used by Behrens and colleagues (Behrens, Hunt, Woolrich, & Rushworth, 2008) (see Figure 2), which has been used in a similar form in previous studies (Diaconescu et al., 2014, 2017).

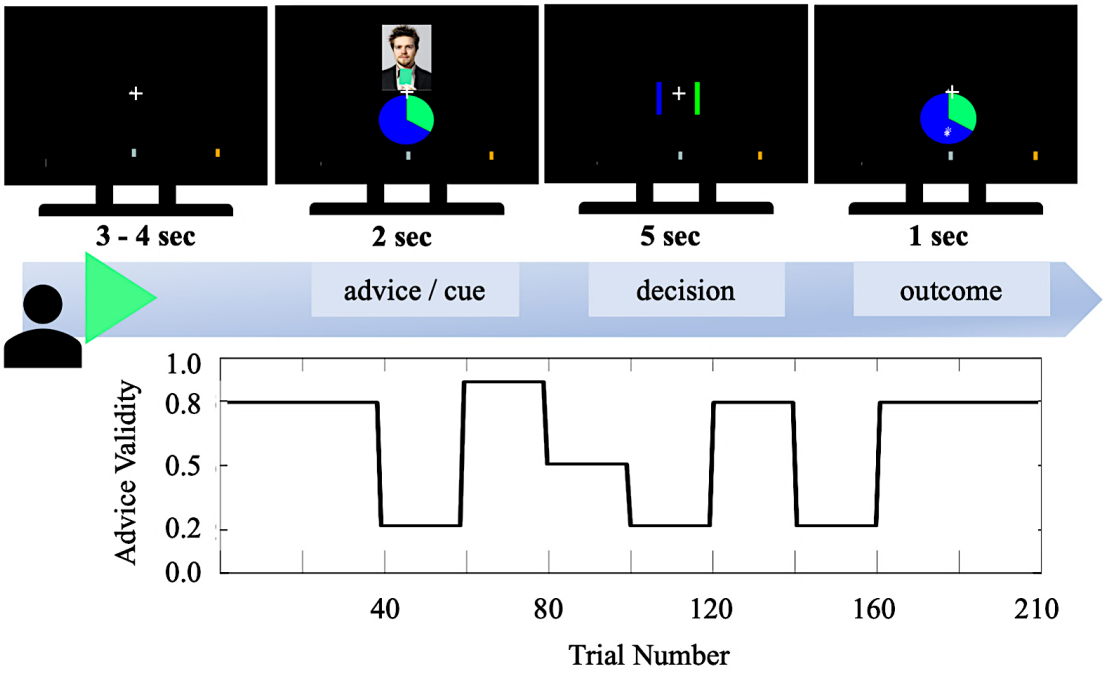

Figure 2 | **Plot of task structure and trajectory of advice validity.**

Participants made binary decisions (blue vs. green) based on a social and a non-social cue (advice and pie-chart) presented simultaneously. Correct predictions result in the accumulation of points growing a progress-bar toward the targets displayed. Surpassing the targets earned participants additional CHF 10 (silver) or 20 (gold). After predicting the 'winning' colour they were informed regarding the real outcome.

Pie-chart probabilities varied between 50:50, 55:45, 60:40, and 75:25.

Advice validity varied across 210 trials (boxcar-chart): the first 42 trials consisted of predominantly (p=0.8) correct advice (1st stable helpful phase), followed by 126 trials of highly variable advice validity (volatile phase). In the last 42 trials, advice was highly valid with p=0.8 again (2nd stable helpful phase). The sequence of trials was fixed and identical across all participants.

Participants played a probabilistic lottery for monetary rewards trying to predict a binary outcome (blue or green) trial by trial (210 trials). Two sources of information, a social and a non-social cue, were presented on each trial. The latter was a pie-chart displaying a veridical probability distribution indicating what colour was more likely to win on any given trial. The pie-chart displayed different green-blue ratios (50:50, 55:45, 60:40, and 75:25) on each trial thus variying in its predictive uncertainty. The social cue (the advice) was represented by a videotaped adviser (recorded in a previous study (Diaconescu et al., 2014)) who gave a recommendation on which colour to choose, by holding up a card (blue or green). All participants were truthfully informed that the adviser did not have full information – and thus could make errors unintentionally – and that each piece of advice had been videotaped in a prior study with the same task (Diaconescu et al., 2014). In the Diaconescu et al (2014) study, an additional control task which was matched in terms of volatility was conducted, in order to ensure that

the main task truly captured social learning (learning from changing intetions). In the control task participants acting as advisers chose the green or blue cards (their "advice") from a card deck which was accurate in either 80% or 20% of the trials. "Advisers" were blind-folded. This allowed the authors to control for the social aspect of the task, while matching the statistical structure of both pie chart information and cue-outcome relations identical. Performance in the control task differed significantly from performance in the main task, even when the card deck was 80% accurate, suggesting that an adviser holding up the cards in a seemingly intentional way was processed differently, which was also captured by a learning model incorporating a social-bias parameter (for more information see (Diaconescu et al., 2014)). Learning in this task should thus be driven by social inference about the adviser's intentions and not represent pure learning of statistics structure. We used one female and one male adviser, and adviser sex was balanced across groups and conditions.

Participants had to make predictions by integrating the two cues. After each decision, they received feedback on their choice and the correct outcome. Players accrued points with every correct prediction. By achieving a cumulative score exceeding the silver or gold targets, participants earned an additional bonus, amounting to approx. 1/6 (silver) or 1/3 (gold) of the experiment's base reimbursement.

The task contained phases of differing advice validity; it began and ended with periods of high advice validity (p=0.80, 42 trials each) and an intermediate period (126 trials) with changes in the advice-outcome contingency (volatility). The trial structure and sequence were fixed and identical across all participants.

2.3.4 Experimental conditions

The two experimental conditions (attributional frames) differed in how potentially unhelpful advice was framed: dispositional (caused by the adviser) vs. situational (caused by the rules of the game). Critically, neither of the frames provided false information but summarised the adviser's role from different angles.

In the dispositional frame, participants' attention was directed to the adviser as a potential source of variability in advice validity, emphasizing his/her ability of acting intentionally in order to achieve his/her own (unknown) goals. In the situational frame, attention was directed to the role of the adviser as part of the task, highlighting that he/she was instructed to use the information available to him/her for guiding the player's behaviour. We induced the two frames by (i) one sentence in the instructions that differed between the two frames, (ii) a reminder on the task start-screen, and (iii) the wording used regarding advice validity. For more details, see the supplementary material.

4. Results

4.1 Sample characteristics

151 of the 162 participants who took part in the experiment scored outside the ±0.5 *sd* intervals of the PCL Frequency scale over all three questionnaire assessments (pre-screening, screening, experiment day), which was the cut-off for group assignment, and were thus eligible for analysis. They were on average 28 years old and had 15.8 years of education (Table 1).

Experimental frame	**Low PD group**		**High PD group**	
Dispositional	*M (SD)*	*range*	*M (SD)*	*range*
age	29.42 *(9.65)*	18-67	26.47 *(8.29)*	18-49
education (years)	17.12 *(3.70)**	11-26	15.31 *(3.11)**	9-21
N	41 (15 male)		36 (15 male)	
Situational				
age	28.8 *(9.74)*	18-54	28.15 *(11.03)*	18-56
education (years)	15.71 *(4.22)*	8-26	14.97 *(3.86)*	7-23
N	40 (17 male)		34 (15 male)	

Table 1 | **Descriptive statistics of participants eligible for analyses.**

N=151, no differences in demographic variables detected between groups.
Groups are not of equal size due to drop-outs (not responding to the invitation or no-shows) and participants being excluded from analyses after data acquisition based on previously defined criteria (see supplementary material).

$^{*}p<0.05$, two-tailed *t*-tests, does not survive Bonferroni correction.
All other variables: $p>0.05$, two-tailed *t*-tests.

Sixty-four of 151 participants did not have an academic background. As expected, participant groups differed across all subscales of the PCL (see Table 2).

	Low PD group		**High PD group**	
	M (SD)	*range*	*M (SD)*	*range*
PCL				
Frequency***	0.33 *(0.34)*	0-2	19.06 *(5.43)*	11-33
Conviction***	0.92 *(1.57)*	0-11	20.44 *(5.77)*	6-35
Distress***	17.88 *(14.80)*	0-46	27.96 *(7.86)*	11-44
BNA Numbers-letters test	15.09 *(3.49)*	6-21	15.33 *(3.43)*	8-21
BNA Symbols test	86.64 *(14.10)*	59-133	83.76 *(15.41)*	60-121

Table 2 | **Questionnaire scores of participants eligible for analyses.**

N=151. Average PCL scores across the three questionnaire assessments.
PCL subscales differed between groups: $^{***}p<0.001$, two-tailed *t*-test, unequal variances, Bonferroni corrected.
Regarding BNA scores (cognitive performance) no differences were detected between groups, $p>0.23$.

Furthermore, no significant difference between groups was detected regarding the cognitive screening administered (BNA working memory: t=0.43, p=0.67; BNA coding: t=-1.20, p=0.23). It is thus unlikely that differences in task behaviour could stem from differences in cognitive capacity.

Participants also did not differ significantly in terms of performance accuracy on the task (two-tailed t-test, df=149, t=-1.82, p=0.07) with high PD participants' accuracy averaging at 0.60 and low PD participants' at 0.61. Furthermore, both groups achieved a similar amount of points (high PD participants earned 41 and low PD participants 47 on average; two-tailed t-test, df=149, t=-1.74, p=0.08) and reached the silver target on average.

4.2 Hypotheses

All following results represent main outcomes of hypothesis tests that were defined in a time-stamped analysis plan prior to data analysis (Analysis plan: https://gitlab.ethz.ch/sibak/sibak-analysis-plan, data: https://www.research-collection.ethz.ch/handle/20.500.11850/333102, and code: https://gitlab.ethz.ch/sibak/sibak-behavior-paper).

Additional effects (both significant and nonsignificant ones) of secondary importance are reported in the supplementary material (see Supplementary Figure 1 for a summary of all hypotheses and results). This paper only reports analyses relating to task-behaviour (Hypotheses I-VI) and the "non-model based approach" in the analysis plan; for corresponding computational analyses, see (Diaconescu et al., 2019).

In addition to inferential statistics (parametric or non-parametric tests, depending on the distribution of the data), we report effect size estimates ($\eta^2_{partial}$ for multiple regression analyses, Cohen's d for t-tests, and Cohen's r for non-parametric tests; (Fritz, Morris, & Richler, 2012)).

4.2.1 Hypothesis I: High PD participants take information provided by the frame less into account than low PD participants *(H_0 rejected)*

We hypothesized that individuals with tendencies to PD behave less adaptively toward differences in social information as they rely on rigid high-level beliefs. We thus expected high PD participants to be less sensitive to the framing effect than low PD participants. We applied a two-way ANOVA with an interaction term (group, frame, and group x frame) to the participants' choices, i.e., how often they chose the colour recommended by the adviser. This group-by-frame interaction was significant regarding participants choosing according to advice overall (df=(1,150), F=5.77, p=0.018, $\eta^2_{partial}$=0.04). Specifically, the interaction plot in Figure 3a suggests that low PD participants more readily made decisions in accordance with the advice under the situational frame (highlighting incorrect advice as circumstantial) than under the dispositional frame (emphasizing incorrect advice as potentially caused by the adviser's intentions).

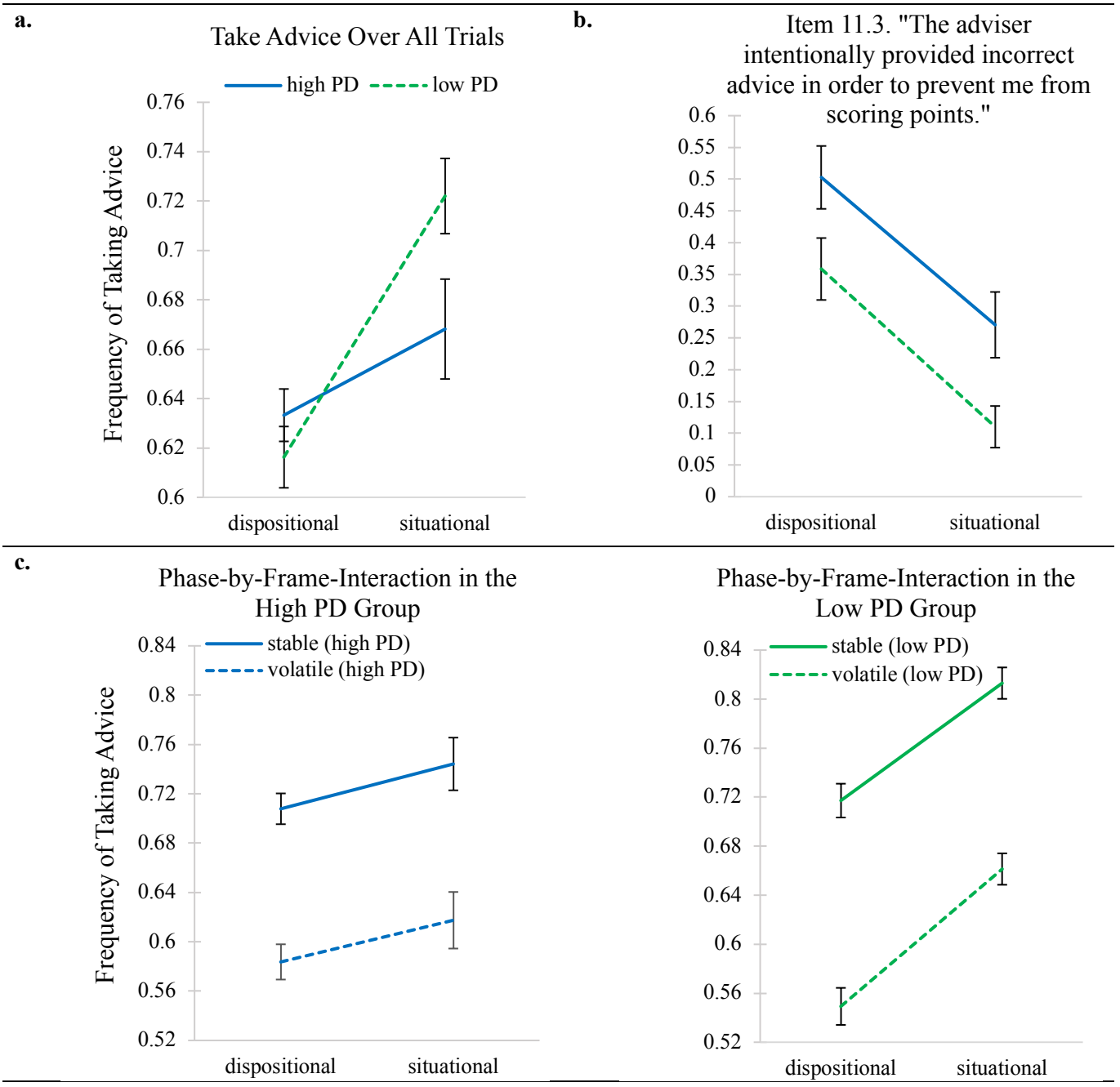

Figure 3 | **ANOVAs for advice-taking behaviour ANOVAs.**

While no differences in advice-taking between frames were identified in high PD participants, low PD participants took advice into account less frequently in the dispositional vs. the situational frame.

a. Two-way interaction for taking advice overall: group*frame: F=5.77, p=0.018, $\eta^2_{partial}$=0.04
frame: F=22.71, p=4.4754e-06, $\eta^2_{partial}$=0.13

b. Main effects regarding agreement with statement:
"*The adviser intentionally provided incorrect advice in order to prevent me from scoring points*" (H. VI)
group: F=11.01, p=0.001, $\eta^2_{partual}$=0.07
frame: F=27.39, p=5.63e-07, $\eta^2_{partial}$=0.16

c. Three-way interaction for taking advice in stable vs. volatile phases: phase*group*frame: F=0.29, p=0.59, $\eta^2_{partial}$=0.002
phase*frame: F=0.15, p=0.7, $\eta^2_{partial}$=0.001
phase*group: F=3.86, p=0.051, $\eta^2_{partial}$=0.03

Y-axes show percentages, error bars indicate standard errors of the mean.

A post-hoc two-tailed *t*-test demonstrated that this difference between frames was significant in the low PD group (df=79, t=-5.55, $p_{Bonferroni}$=7.3e-07, *Cohen's d*=0.36). In contrast, no significant difference in advice-taking behaviour between framing conditions could be detected in high PD participants (df=68, t=-1.52 p=0.13, *Cohen's d*=0.11, suggesting that they did not integrate and utilize the information provided by the frame.

4.2.2 Hypothesis II: High PD participants attribute misleading advice validity to the adviser's character rather than to the possibility of the adviser making a mistake due to incomplete information. *(H_0 not rejected)*

We tested this hypothesis with regard to participants' advice-taking in the task phases where advice was stable and helpful vs. when advice validity was highly variable. We predicted that in the situational frame, high PD participants would more frequently choose against the advice when it was volatile, and that this frame-specific difference between volatile and stable phases would be stronger than in low PD participants. In the dispositional frame, we expected smaller group differences. However, this three-way interaction was not significant (phase*group*frame: F=0.29, p=0.59, $\eta^2_{partial}$=0.002, see Figure 3c and the Supplementary Material for more information and analyses).

4.2.3 Hypothesis III: High PD compared to low PD participants generally attribute negative events to more external-personal causes. *(H_0 not rejected)*

We tested this with debriefing question 14 *("In your opinion, what factors determined your performance in the task?",* participants distributed a total of 100% to the following answer options: *"*a. *You as the player"*, *"*b. *The adviser which was appointed to you"*, and *"*c. *The rules / the structure of the game"*). This hypothesis was not confirmed (see supplementary material for more information).

4.2.4 Hypothesis IV: High PD participants attribute differences in advice validity to the adviser being malevolent. *(H_0 rejected)*

We hypothesized that high PD participants exhibited more negative beliefs about the adviser and more readily attributed differences in advice validity to the adviser's character than low PD individuals and that this effect would be moderated by the framing. We tested this with debriefing item "*The adviser intentionally gave false information, without profiting from it.*", which states a possible cause for incorrect advice. Participants had to indicate how much they believed this cause played a role for the adviser providing incorrect advice (percentage assignment 0%-100%). While the interaction was not significant, the main effect of group was (df=(1,150), F=16.87 p=6.63e-05, $\eta^2_{partial}$=0.1). On average, high PD participants assigned 38.6±29.2 % to this statement whereas low PD participants assigned 20.3±25.9 %.

In the debriefing questionnaire, participants were asked to rate what caused incorrect advice, and assign percentages to the adviser and to the rules of the game (debriefing question 13: "*When the adviser provided you with misleading/incorrect advice: What do you believe was the cause of this?".* Participants distributed a total of 100% to the following answer options: *"*a. *The adviser which was appointed to you"* and *"*b. *The rules/the structure of the game"*. Given the nature of the distributions

(percentages assigned to the different answer options were tied because they had to sum up to 100%), we used a Wilcoxon rank sum test to compare groups. As expected, high PD participants were more likely to rate the adviser as the cause for incorrect advice (40.0±23.8%), compared to low PD participants (29.6±27.7%; z=2.42, p=0.008, *Cohen's r*=0.2).

4.2.5 Hypothesis V: High PD participants expect receiving misleading advice. *(H_0 not rejected)*

We tested this hypothesis with the debriefing questionnaire (administered after the task), where we asked what advice participants expected to receive before playing the game (*1=correct advice, 6=incorrect advice*). We expected an interaction effect with high PD scoring higher than low PD participants in the situational frame but not the dispositional frame. While this interaction did not reach significance, both main effects did, showing that low PD participants scored lower on this scale on average (2.81±1.18), indicating a tendency toward expecting correct advice compared to high PD participants (3.45±1.19) than low PD participants (main effect of group; df=147, F=11.72, p=8.0e-04, $\eta^2_{partial}$=0.07, see supplementary for full statistics).

4.2.6 Hypothesis VI: High PD participants view incorrect advice as directed towards them. *(H_0 not rejected)*

This could be directly assessed via debriefing questionnaire item *"The adviser intentionally provided incorrect advice in order to prevent me from scoring points."* (percentage assignment 0%-100%). We originally expected an interaction here, with low PD participants assigning higher percentages to this statement in the dispositional compared to the situational frame and high PD participants assigning high percentages across framing conditions. However, only the main effect of group reached significance (df=(1,150), F=11.01, p=0.001, $\eta^2_{partial}$=0.07), with high PD participants endorsing the statement more (39±31.9%) than low PD participants (23.6±29.2%, see Figure 3b).

5. Discussion

In this study, we investigated how individuals scoring high vs. low on the Paranoia Checklist (PCL; (Freeman et al., 2005)) incorporate attributional priors into learning from advice. Participants performed a probabilistic advice-taking paradigm with variable advice-outcome contingencies under one of two experimental frames which differentially emphasised causes of social information (dispositional vs. situational attributional frames).

We found that low PCL scorers took advice into account less under the dispositional (highlighting the adviser as the cause for incorrect advice) vs. the situational frame (highlighting incorrect advice as circumstantial), whereas high PCL scorers did not differ between framing conditions (Figure 3a). High PD participants' behaviour is similar to what was reported in a recent study using the "dictator game" (Raihani & Bell, 2017); the authors induced two differing experimental contexts (being at the receiving end of a dictator's decisions or a third-person observer) and found that persecutory ideation was related to attributing harmful intent, irrespective of context (Raihani & Bell, 2017). Low PD participants' advice-taking behaviour corresponds to what Fouragnan et al. found in a recent study, namely that prior knowledge about an agent's reputation generally influences learning about intentions (Fouragnan et al., 2013). Specifically, healthy participants relied on experimentally-induced reputation priors in a social learning task, even in light of disconfirming evidence.

Our dispositional frame, introduced a negative prior about the adviser's intentions, prompting low PD participants to disregard advice more often. The group-by-frame interaction (Figure 3a) indicates that low scorers' advice-taking behaviour seemed to be driven by these experimentally-induced priors. High scorers however disregarded the advice irrespective of experimental frame, even in the "safer" social context when the rules of the game were emphasised as causes for incorrect advice.

One explanation for high PD participants' lack of differences in advice-taking behaviour across frames, might be that – rather than the experimentally-induced priors – their high-level prior beliefs about the adviser's intentions influenced their decisions. Specifically, from the perspective of hierarchical models of Bayesian inference, when low-level beliefs (e.g. about trial-by-trial behaviour) are ambiguous (have high variance) and higher-level beliefs (e.g. about the intentions of others) are held with more conviction (precision), perception and learning will be more strongly influenced by higher-level beliefs. Thus, the reduced impact of the framing in this study might reflect high PD participants' reliance on strong (precise) higher-level prior beliefs in the face of weaker (less precise) experimentally-induced, lower-level predictions. This explanation is in line with findings showing an increased influence of prior beliefs on the perception of ambiguous stimuli (Schmack et al., 2013) as well as reduced use of experimentally-induced priors in delusion-prone individuals (Stuke, Weilnhammer, Sterzer, & Schmack, 2018), suggesting that delusion(-proneness) is characterised by differences in the precision of prior beliefs at different levels of the processing hierarchy.

Indeed, high PD participants reported viewing the adviser as the main cause of incorrect advice, as opposed to considering the adviser having incomplete information (e.g., Hypothesis 4, debriefing question 13 "*When the adviser provided you with misleading/incorrect advice: What do you believe was the cause of this?",* high PD participants agreed more with *"The adviser which was appointed to you"*). Furthermore, compared to low PD participants, they reported expecting misleading advice more (Hypothesis 5) and viewing the adviser as acting intentionally malevolent towards them more (Hypothesis 6).

These findings suggest that high PCL scorers relied on overly precise negative higher-level prior beliefs when inferring on the adviser's intentions Following a reviewer's suggestion to scrutinise this interpretation in additional exploratory analyses, we investigated whether participants' conviction scores on the PCL (obtained prior to the experiment) were related to how they answered debriefing questionnaire item 11.3 ("*The adviser intentionally provided incorrect advice in order to prevent me from scoring points*."). We found that the higher the conviction ratings regarding persecutory beliefs as assessed with the PCL, the more the participant judged the adviser as acting intentionally and malevolently (F=12.8, p=0.0005, $\eta^2_{partial}$=0.07).

An alternative interpretation for the lack of between-condition differences in high PD participants might be that they did not believe the experimental framing that was induced via task instruction. We do not think this was the case, for the following reasons:

(i) We asked participants at the end of the debriefing questionnaire for feedback on different aspects of the study (debriefing questions 17 and 18), whether the instructions were understandable (Q 19), and if they felt influenced by them in how they viewed the adviser (Q 20) and how they played the game (Q 21). None of the participants mentioned suspecting anything or being influenced. Furthermore, after having been debriefed about the study and the framing, participants stated not having suspected anything.

(ii) Additionally, in the debriefing questionnaire, we asked participants to indicate their agreement with two statements probing the information highlighted in the two experimental frames: The adviser not having full information being the cause of incorrect advice in the situational frame (Q 11.5, „The adviser, in general, did not have full information and made mistakes.") and the possibility of the adviser being the cause of incorrect advice in the dispositional frame (Q 11.4, „The adviser intentionally provided false information because this was part of his/her instructions/task.").

We found a significant main effect of framing, together with non-significant main effect of group and interactions (see Supplementary Table 3 for detailed statistics). This suggests that (i) both high and low PD participants viewed the adviser's incomplete information as a more likely cause for incorrect advice in the situational compared to the dispositional frame (Q 11.5) and that the adviser providing false information due to their instructions was the more likely cause for incorrect advice in the dispositional frame compared to the situational frame (Q 11.4). These effects were expected based on the framing

instructions. Thus, participants generally seemed to have a similar understanding of the framing and did not disregard the framing-specific instructions (see 8.2.1 and 8.2.1 in the supplementary material for detailed statistics).

Our findings align with previous reports of theory of mind (ToM) deficits in psychosis patients with delusions, indicating an external-personal attribution bias (Bliksted, Fagerlund, Weed, Frith, & Videbech, 2014; Craig et al., 2004; Frith & Corcoran, 1996; Langdon, Coltheart, Ward, & Catts, 2001; Langdon et al., 2006), and extend these findings to subclinical PD. The association between social cognition and PD in subclinical populations has thus far been inconclusive ((McKay et al., 2005) and for a review see (Garety & Freeman, 2013)), potentially due to ToM paradigms being predominantly questionnaire-based measures. Although recent investigations of external-attributions in persecutory ideation captured dynamic aspects of social inference (Raihani & Bell, 2017), how predictions are updated as a result of contradicting evidence or PEs when information is continuously changing has not been examined yet.

The results of this study suggest that individuals with PD tendencies may incorporate PEs less into their predictions about the adviser's intentions due to more rigid high-level prior beliefs about the adviser's intentions. In a separate analysis, computational modelling of our data suggested that maladaptive social inference in the task may result from overly precise higher-level beliefs about the adviser's fidelity, leading to reduced belief-updating in the face of incoming PEs (Diaconescu et al., 2019).

In this study, we aimed to capture the behaviour of individuals for whom persecutory ideation was a stable trait. Since impairments of social cognition might contribute to risk for developing psychosis and are found in first-episode psychosis (FEP) patients (Bora & Pantelis, 2013), (Sun et al., 2011), future extension of our approach might serve to address clinically-relevant predictions, such as transition to psychosis in clinical high risk individuals and treatment response in FEP patients.

6. Acknowledgements

This work was supported by the René and Susanne Braginsky Foundation (KES), the University of Zurich (KES), and the Swiss National Science Foundation Ambizione (PZ00P3_167952 to AOD).

We would also like to thank Dr. Fabien Vinckier for providing very helpful comments.

7. References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The Computational Anatomy of Psychosis. *Frontiers in Psychiatry*, *4*(May), 1–26. https://doi.org/10.3389/fpsyt.2013.00047

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538.Associative

Bell, V., & Halligan, P. W. (2013). The neural basis of abnormal personal belief. In *The Neural Basis of Human Belief Systems* (pp. 191–224).

Biedermann, F., Frajo-Apor, B., & Hofer, A. (2012). Theory of mind and its relevance in schizophrenia. *Current Opinion in Psychiatry*, *25*(2), 71–75. https://doi.org/10.1097/YCO.0b013e3283503624

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., & Videbech, P. (2014). Social cognition and neurocognitive deficits in first-episode schizophrenia. *Schizophrenia Research*, *153*(1–3), 9–17. https://doi.org/10.1016/j.schres.2014.01.010

Bora, E., & Pantelis, C. (2013). Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: Systematic review and meta-analysis. *Schizophrenia Research*, *144*(1–3), 31–36. https://doi.org/10.1016/j.schres.2012.12.013

Coltheart, M., Menzies, P., & Sutton, J. (2010). Abductive inference and delusional belief. *Cognitive Neuropsychiatry*, *15*(1–3), 261–287. https://doi.org/10.1080/13546800903439120

Corlett, P. R., Honey, G. D., & Fletcher, P. C. (2016). Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology*, *30*(11), 1145–1155. https://doi.org/10.1177/0269881116650087

Corlett, P. R., Taylor, J. R., Wang, X. J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, *92*(3), 345–369. https://doi.org/10.1016/j.pneurobio.2010.06.007

Craig, J. S., Hatton, C., Craig, F. B., & Bentall, R. P. (2004). Persecutory beliefs, attributions and theory of mind: Comparison of patients with paranoid delusions, Asperger's syndrome and healthy controls. *Schizophrenia Research*, *69*(1), 29–33. https://doi.org/10.1016/S0920-9964(03)00154-3

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., … Stephan, K. E. (2014). Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLoS Computational Biology*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(November 2016), 1–17. https://doi.org/10.1093/scan/nsw171

Diaconescu, A. O., Wellstein, K. V., Mathys, C., & Stephan, K. E. (2019). Hierarchical Bayesian models of social inference for probing persecutory delusional ideation. *Journal of Abnormal Psychology*, *under revi*(Predictive Coding and Psychopathology).

Ermakova, A. O., Gileadi, N., Knolle, F., Diaz, A. J., & Anderson, R. (2017). Cost evaluation during decision making in patients at early stages of psychosis. *Computational*

Psychiatry, *in press*, 1–39. https://doi.org/10.1101/225920

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fervaha, G., Hill, C., Agid, O., Takeuchi, H., Foussias, G., Siddiqui, I., … Remington, G. (2014). Examination of the validity of the Brief Neurocognitive Assessment (BNA) for schizophrenia. *Schizophrenia Research*, *166*(1–3), 304–309. https://doi.org/10.1016/j.schres.2015.05.015

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58. https://doi.org/10.1038/nrn2536

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational Priors Magnify Striatal Responses to Violations of Trust. *Journal of Neuroscience*, *33*(8), 3602–3611. https://doi.org/10.1523/JNEUROSCI.3086-12.2013

Freeman, D. (2007). Suspicious minds: The psychology of persecutory delusions. *Clinical Psychology Review*, *27*(4), 425–457. https://doi.org/10.1016/j.cpr.2006.10.004

Freeman, D., & Garety, P. (2014). Advances in understanding and treating persecutory delusions: A review. *Social Psychiatry and Psychiatric Epidemiology*, *49*(8), 1179–1189. https://doi.org/10.1007/s00127-014-0928-7

Freeman, D., Garety, P. A., Bebbington, P. E., Smith, B., Rollinson, R., Kuipers, E., … Katarzyna, R. (2005). Psychological investigation of the structure of paranoia in a non-clinical population. *British Journal of Psychiatry*, *186*, 427–435. https://doi.org/10.1192/bjp.186.5.427

Freeman, D., Startup, H., Dunn, G., Wingham, G., Černis, E., Evans, N., … Kingdon, D. (2014). Persecutory delusions and psychological well-being. *Social Psychiatry and Psychiatric Epidemiology*, *49*(7), 1045–1050. https://doi.org/10.1007/s00127-013-0803-y

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Frith, C. D., & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine*, *26*(03), 521. https://doi.org/10.1017/S0033291700035601

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Garety, P. A., & Freeman, D. (2013). The past and future of delusions research: From the inexplicable to the treatable. *British Journal of Psychiatry*, *203*(5), 327–333. https://doi.org/10.1192/bjp.bp.113.126953

Garety, P. A., Hemsley, D. R., & Wessley, M. R. C. (1991). Reasoning in Deluded Schizophrenic and Paranoid Patients. Biases in Performance on a Probabilistic Inference Task. *The Journal of Nervous and Mental Disease*, *179*(4), 149–201.

Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, *57*(6), 940–949. https://doi.org/10.1037/0022-3514.57.6.940

Heinz, a. (2002). Dopaminergic dysfunction in alcoholism and schizophrenia--

psychopathological and behavioral correlates. *European Psychiatry : The Journal of the Association of European Psychiatrists*, *17*(1), 9–16. https://doi.org/S0924933802006284 [pii]

Hemsley, D. R., & Garety, P. A. (1986). The Formation of Maintenance of Delusions : *The British Journal of Psychiatry*, *149*(1), 51–56. https://doi.org/10.1192/bjp.149.1.51

Kapur, S. (2003). Psychosis as a State of Aberrant Salience : and Pharmacology in Schizophrenia. *American Journal of Psychiatry*, *160*, 13–23. https://doi.org/10.1176/appi.ajp.160.1.13

Keers, R., Ullrich, S., DeStavola, B. L., & Coid, J. W. (2014). Association of violence with emergence of persecutory delusions in untreated schizophrenia. *American Journal of Psychiatry*, *171*(3), 332–339. https://doi.org/10.1176/appi.ajp.2013.13010134

Kinderman, P., & Bentall, R. P. (1996). A new measure of causal locus: The internal, personal and situational attributions questionnaire. *Personality and Individual Differences*, *20*(2), 261–264. https://doi.org/10.1016/0191-8869(95)00186-7

Langdon, R., Coltheart, M., Ward, P. B., & Catts, S. V. (2001). Mentalising, executive planning and disengagement in schizophrenia. *Cognitive Neuropsychiatry*, *6*(2), 81–108. https://doi.org/10.1080/13546800042000061

Langdon, R., Corner, T., McLaren, J., Ward, P. B., & Coltheart, M. (2006). Externalizing and personalizing biases in persecutory delusions: The relationship with poor insight and theory-of-mind. *Behaviour Research and Therapy*, *44*(5), 699–713. https://doi.org/10.1016/j.brat.2005.03.012

Lincoln, T. M., Mehl, S., Exner, C., Lindenmeyer, J., & Rief, W. (2010). Attributional style and persecutory delusions. Evidence for an event independent and state specific external-personal attribution bias for social situations. *Cognitive Therapy and Research*, *34*(3), 297–302. https://doi.org/10.1007/s10608-009-9284-4

Lincoln, T. M., Ziegler, M., Lüllmann, E., Müller, M. J., & Rief, W. (2010). Can delusions be self-assessed? Concordance between self- and observer-rated delusions in schizophrenia. *Psychiatry Research*, *178*(2), 249–254. https://doi.org/10.1016/j.psychres.2009.04.019

Martin, J. A., & Penn, D. L. (2002). Attributional Style in Schizophrenia: An Investigation in Outpatients With and Without Persecutory Delusions. *Schizophrenia Bulletin*, *28*(1), 131–141. https://doi.org/10.1093/oxfordjournals.schbul.a006916

McCrae, R. R., & Costa, P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, *36*(3), 587–596. https://doi.org/10.1016/S0191-8869(03)00118-1

McKay, R., Langdon, R., & Coltheart, M. (2005). Paranoia, persecutory delusions and attributional biases. *Psychiatry Research*, *136*(2–3), 233–245. https://doi.org/10.1016/j.psychres.2005.06.004

Moutoussis, M., Bentall, R. P., El-Deredy, W., & Dayan, P. (2011). Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. *Cognitive Neuropsychiatry*, *16*(5), 422–447. https://doi.org/10.1080/13546805.2010.548678

Peters, E., & Garety, P. (2006). Cognitive functioning in delusions: A longitudinal analysis. *Behaviour Research and Therapy*, *44*(4), 481–514. https://doi.org/10.1016/j.brat.2005.03.008

Raihani, N. J., & Bell, V. (2017). Paranoia and the social representation of others: A large-scale game theory approach. *Scientific Reports*, *7*(1), 1–9.

https://doi.org/10.1038/s41598-017-04805-3

Raihani, N. J., & Bell, V. (2018). Conflict and cooperation in paranoia: A large-scale behavioural experiment. *Psychological Medicine*, *48*(9), 1523–1531. https://doi.org/10.1017/S0033291717003075

Rao, R. P. N., & Ballard, D. H. (1999). Predictive Coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Sartorius, N., Sartorius, N., Jablensky, A., Jablensky, A., Korten, A., Korten, A., … Day, R. (1986). Early Manifestations and First-Contact Incidence of Schizophrenia in Different Cultures. *Psychological Medicine*, *16*, 909–928.

Schmack, K., Gomez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rossler, H., Haynes, J.-D., … Sterzer, P. (2013). Delusions and the Role of Beliefs in Perceptual Inference. *Journal of Neuroscience*, *33*(34), 13701–13712. https://doi.org/10.1523/JNEUROSCI.1778-13.2013

Shaner, A. (1999). Delusions, superstitious conditioning and chaotic dopamine neurodynamics. *Medical Hypotheses*, *52*(2), 119–123. https://doi.org/10.1054/mehy.1997.0656

So, S. H., Freeman, D., Dunn, G., Kapur, S., Kuipers, E., Bebbington, P., … Garety, P. A. (2012). Jumping to conclusions, a lack of belief flexibility and delusional conviction in psychosis: A longitudinal investigation of the structure, frequency, and relatedness of reasoning biases. *Journal of Abnormal Psychology*, *121*(1), 129–139. https://doi.org/10.1037/a0025297

Speechley, W. J., Whitman, J. C., & Woodward, T. S. (2010). The contribution of hypersalience to the "jumping to conclusions" bias associated with delusions in schizophrenia. *Journal of Psychiatry and Neuroscience*, *35*(1), 7–17. https://doi.org/10.1503/jpn.090025

Stefanis, N. C., Hanssen, M., Smirnis, N. K., Avramopoulkos, D. A., Evdokimidis, I. K., Stefanis, C. N., … Van Os, J. (2002). Evidence that three dimensions of psychosis have a distribution in the general population. *Psychological Medicine*, *32*, 347–358.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., … Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2018.05.015

Stuke, H., Weilnhammer, V. A., Sterzer, P., & Schmack, K. (2018). Delusion Proneness is Linked to a Reduced Usage of Prior Beliefs in Perceptual Decisions. *Schizophrenia Bulletin*, 1–7. https://doi.org/10.1093/schbul/sbx189

Sun, H., Young, N., Hwan, J., Kim, E., Shim, G., Yoon, H., … Soo, J. (2011). Social cognition and neurocognition as predictors of conversion to psychosis in individuals at ultra-high risk. *Schizophrenia Research*, *130*(1–3), 170–175. https://doi.org/10.1016/j.schres.2011.04.023

Van Os, J., Verdoux, H., Maurice-Tison, S., Gay, B., Liraud, F., Salamon, R., & Bourgeois, M. (1999). Self-reported psychosis-like symptoms and the continuum of psychosis. *Social Psychiatry and Psychiatric Epidemiology*, *34*(9), 459–463. https://doi.org/10.1007/s001270050220

Ventura, J., Wood, R. C., & Hellemann, G. S. (2013). Symptom domains and neurocognitive functioning can help differentiate social cognitive processes in schizophrenia: A meta-analysis. *Schizophrenia Bulletin*, *39*(1), 102–111. https://doi.org/10.1093/schbul/sbr067

Inflexible social inference in individuals with subclinical persecutory delusional tendencies

8. Supplementary material

8.1. Methods

The instructions and debriefing questionnaire, which were originally presented to participants in their native German, were translated into English for the purpose of this paper.

8.1.1 Framing Conditions

The framing was supported via two different channels: (i) one sentence in the task instructions and (ii) a start screen. Pronouns were adapted to the adviser's gender.

i) Task instructions

Dispositional Frame
"… The adviser has generally more information than you about the outcome on each trial. **The objective of the adviser is to use this information to reach his/her own goals.** Note that the adviser does not have 100% accurate information about which colour "wins" and he/she might be incorrect. Nevertheless, he/she will on average have better information than you and his/her advice may be valuable to you. …"

Situational Frame
"… The adviser has has generally more information than you about the outcome on each trial. **The objective of the adviser is to use this information to guide your choices**. Note that the adviser does not have 100% accurate information about which colour "wins" and he/she might be incorrect. Nevertheless, he/she will on average have better information than you and his/her advice may be valuable to you.. …"

ii) Start screens

The start screen of the social learning task was presented for 8 seconds before the task started. It included a short reminder regarding the adviser, which supported the respective framing. In the dispositional frame, a picture of the adviser the participant would interact with was presented, whereas in the situational frame the image of the task structure (as seen previously by all participants on the paper instructions) was presented. See the reminder sentences for each framing condition below. Pronouns were adapted to the adviser's gender:

Dispositional Frame (case of male adviser)
"You are playing the "Social Learning" game with this adviser.
Do not forget, that the adviser has his own goals and could sometimes try to mislead you."

Situational Frame (case of female adviser)
"You are playing the "Social Learning" game in the role of the player.
Do not forget that the the adviser does not have full information. For this reason, she might mistakenly give you incorrect advice even when she would like to help you."

iii) Wording

In the debriefing questionnaire as well as in the instructions the following wording differed between frames:

Dispositional Frame

- valid advice was worded as “helpful advice”
- invalid advice was worded as “misleading advice”

Situational Frame

- valid advice was worded as “correct advice”
- invalid advice was worded as “incorrect advice”

8.1.2. Debriefing Questionnaire

The Debriefing Questionnaire contained 16 questions, only those relevant to the analyses discussed in this paper are presented here:

02. How happy are you personally with your performance on this task?

not at all happy — moderately happy — very happy

0----------0----------0----------0----------0----------0

10. After the practice round, what expectations did you have with regard to the adviser’s behaviour in the upcoming task? Did you expect him/her to give you rather helpful/correct or misleading/incorrect advice?

helpful/correct advice — misleading/incorrect advice

0----------0----------0----------0----------0----------0

11. The adviser did not always give you helpful/correct advice. What role do you believe the adviser played in this?
Please indicate how much you consider the statements below to be true:

II. The adviser intentionally gave false information, without profiting from it.

0--------0--------0--------0--------0--------0--------0--------0--------0--------0--------0
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

III. The adviser intentionally provided incorrect advice in order to prevent me from scoring points.

0--------0--------0--------0--------0--------0--------0--------0--------0--------0--------0
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

IV. The adviser intentionally provided false information because this was part of his/her instructions/task.

0--------0--------0--------0--------0--------0--------0--------0--------0--------0--------0
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

V. The adviser, in general, did not have full information and made mistakes.

0--------0--------0--------0--------0--------0--------0--------0--------0--------0--------0
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

13. When the adviser provided you with misleading/incorrect advice: What do you believe was the cause of this? *Please assign percentages to the different aspects such that you reach a sum of 100% in total (you can also assign 0%).*[1]

a) The adviser which was appointed to you _____
b) The rules of the game / the structure of the game _____

[1] Note that we asked participants to answer this question regarding three different time-points during the task, i.e. at the beginning, in the middle and at the end of the task. We did that in order to make sure that participants really thought about their beliefs across the entire interaction with the adviser.

14. In your opinion, what factors determined your performance in the task?
Please assign percentages to the different aspects such that you reach a sum of 100% in total (you can also assign 0%).
a) You as the player
b) The adviser which was appointed to you
c) The rules of the game / the structure of the game

15. What impression of the adviser did you have **at the beginning** of the game (first interaction)?
unlikeable likeable
0----------0----------0----------0----------0----------0

16. What impression of the adviser did you have **at the end** of the game (after finishing)?
unlikeable likeable
0----------0----------0----------0----------0----------0

Part C: Regarding the implementation of the experiment

17. Did wearing headphones help your concentration?
o Yes o No If yes, please describe:

18. During the study, did you feel that you were able to ask questions whenever necessary?
o Yes o No If yes, please describe:

19. Were the instructions generally understandable?
o Yes o No If yes, please describe:

20. Did the instructions, which you received regarding the task, influence your jugdment of the adviser?
o Yes o No If yes, please describe:

21. Did the instructions, which you received regarding the task, influence the strategy you employed?
o Yes o No If yes, please describe:

22. Would you like to leave any other comments? More thoughts regarding the task, the experiment or anything else?

8.1.2 Participant exclusion criteria for analyses

In the analysis plan (https://gitlab.ethz.ch/sibak/sibak-analysis-plan) we defined the following conditions for participants to be excluded from analyses:

1) Participants who did not score 0.5 sd below (respectively above) the mean of the PCL Frequency scale after filling in the screening questionnaire for a third time at experiment day were excluded from analysis because they no longer fulfilled the cut-off for the assignment to either the low or the high PD group.
2) Four participants were part of a short pilot experiment of the stimulus input structure, which was then adjusted for the main study. These participants performed the task with a different learning trajectory and will therefore be excluded from behavioural analyses but will still be eligible for questionnaire data analyses.
3) Participants exhibiting performance accuracy of below 50%.
4) Participants who reported having pressed the wrong button on more than 10% of trials.

Ten participants were excluded from analysis due to 1) their PCL scores not meeting group assignment criteria any longer. One participant had to be excluded due to 3) performance accuracy below 0.5.

8.2. Results

In this section, we report additional statistical tests and results, which we were unable to discuss in the main text due to space limitations (see Supplementary Figure 1 for a summary of all hypotheses and results).

Hypotheses	Dependent Variables	Effects
Hypothesis I: High PD participants take information provided by the frame less into account than low PD participants. **Hypothesis II:** High PD participants attribute misleading advice to the adviser's character rather than to the possibility of the adviser making a mistake due to incomplete information. **Hypothesis III:** High PD compared to low PD participants generally attribute negative events to more external-personal causes. **Hypothesis IV:** High PD participants attribute differences in advice validity to the adviser being malevolent. **Hypothesis V:** High PD participants expect receiving misleading advice. **Hypothesis VI:** High PD participants view incorrect advice as directed towards them.	**Choosing in accordance with advice overall** {Hypothesis I.}	**IA effect group*frame ✓ {interaction predicted by H I.}** Main effect group ✗ Main effect frame ✓
	Choosing in accordance with advice when pie chart is at p=0.5 {Hypothesis I.}	**IA effect group*frame ✗ {interaction predicted by H I.}** Main effect group ✗ Main effect frame ✓
	Item 10: expectation of incorrect/correct advice {Hypothesis I. and Hypothesis V.}	**IA effect group*frame ✗ {interaction predicted by H I. & H V.}** Main effect group ✓ Main effect frame ✓
	Item 11.2: Adviser intentionally gave incorrect advice {Hypothesis I.}	**IA effect group*frame ✗ {interaction predicted by H I.}** Main effect group ✓ Main effect frame ✓
	Item 11.3: Adviser intended to prevent from scoring points {Hypothesis I. and Hypothesis VI.}	**IA effect group*frame ✗ {interaction predicted by H I. & H VI.}** Main effect group ✓ Main effect frame ✓
	Item 11.5: Adviser did not have full information {Hypothesis I. and Hypothesis II.}	**IA effect group*frame ✗ {interaction predicted by H I.}** **Main effect group ✗ {main effect predicted by H II.}** Main effect frame ✓
	Item 13: attribution of incorrect advice on adviser vs. rules of the game {Hypothesis I. and Hypothesis IV.}	**IA effect group*frame ✗ {interaction predicted by H I. & H IV.}** **Main effect group ✓ {main effect predicted by H IV.}** Main effect frame ✗
	Item 14: attribution of bad performance on adviser vs. rules of the game {Hypothesis III.}	**IA effect group*frame ✗ {interaction predicted by H III.}** Main effect group ✗ Main effect frame ✗
	Choosing in accordance with advice in volatile vs. stable trials {Hypothesis II.}	**IA effect phase*group*frame ✗ {interaction predicted by H II.}** IA phase*group ✗ IA phase*frame ✗
	Choosing in accordance with advice before vs. after volatility {Hypothesis VI.}	**IA effect phase*group*frame ✗ {interaction predicted by H VI.}** IA phase*group ✗ IA phase*frame ✗

Supplementary Figure 1 | **Overview of hypotheses and tested variables**

8.2.1 Questionnaire scores

We administered the CAPE (*Community Assessment of Psychic Experiences* (Stefanis et al., 2002)) at the end of experiment day and found – as expected – significant group differences on all three dimensions: depressive symptoms, negative symptoms, and positive symptoms.

	Low PD group		**High PD group**	
	M (SD)	*range*	*M (SD)*	*range*
CAPE Depressive dimension ***	2.90 *(0.84)*	0.38-5.38	4.80 *(1.11)*	2.50-7.38
Frequency***	1.61 *(0.27)*	1.13-2.50	2.36 *(0.52)*	1.38-3.75
Distress***	1.28 *(0.64)*	0.00-3.38	2.44 *(0.69)*	0.88-3.75
CAPE Negative dimension***	2.65 *(0.77)*	1.14-4.93	4.29 *(0.92)*	2.43-7.07
Frequency***	1.52*(0.28)*	1.00-2.29	2.21*(0.46)*	1.43-3.64
Distress***	1.12 *(0.57)*	0.00-2.71	2.08 *(0.57)*	0.93-3.43
CAPE Positive dimension***	1.84 *(0.57)*	1.00-4.47	3.14 *(0.75)*	1.58-6.00
Frequency***	1.22 *(0.17)*	1.00-1.74	1.79 *(0.32)*	1.26-2.95
Distress***	0.62 *(0.52)*	0.00-3.42	1.34 *(0.49)*	0.32-3.26

Supplementary Table 1 | **CAPE questionnaire scores for participants eligible for analysis.**

The CAPE is a self-report questionnaire assessing depressive, negative, and positive symptoms in the general population consisting of 42. Subscales include frequency of the thoughts/feelings having occurred in participants lifetime and distress these thoughts caused. On the frequency dimension 1=never and 4=almost always, on the distress dimension participants who had a frequency score =1 on the same item get 0 for the respective item, otherwise 1= not distressed, 4 = very distressed.

N=151*; two-tailed t-tests for unequal variances, Bonferroni corrected, *** p<0.001

8.2.2 Hypothesis I: High PD participants take information provided by the frame less into account than low PD participants

We hypothesised that participants in the high PD group should take situational information less into account than participants in the low PD group. Furthermore, assuming that participants in the high PD group have an *a priori* tendency of assigning malevolent intentions to the adviser, we predicted that the (negative) dispositional framing would affect participants in the high PD group less than those in the low PD group. We thus expected a group-by-frame interaction where the frame influences decisions more in the low PD group than in the high PD group.

8.2.2.1 Discussion of task-behaviour related effects not reported in the main paper

In the analysis plan, we stated that we would test these hypotheses using ANOVAs with interaction effects for two categorical variables (group and frame) with respect to (i) choosing in accordance with advice overall and (ii) choosing in accordance with advice when the pie chart is at chance. In the main text of the paper, we report the significant interaction effects with respect to the first part of the hypothesis, namely choosing in accordance with advice overall; see Supplementary Table 2 for more detailed statistics.

	ANOVA Table for **(i) 'taking advice over all trials'**			
	SumSq	***Df***	***F***	***p***
Framing Condition	0.186	1	22.713	4.475e-06 ***
Group	0.013	1	1.561	0.213
Group x Frame (IA)	0.047	1	5.768	0.018 *
Error (within groups)	1.201	147	-	-
Total	1.461	150	-	-

	ANOVA Table for **(ii)'taking advice when advice validity = 0.5'**			
	SumSq	***Df***	***F***	***p***
Framing Condition	0.150	1	9.645	0.002 **
Group	9.359e-06	1	6.032e-04	0.980
Group x Frame (IA)	0.032	1	2.08	0.151
Error (within groups)	2.281	147	-	-
Total	2.474	150	-	-

Supplementary Table 2 | **Two-way ANOVA with interaction term on task behaviour.**
N=151, significance levels: * p<.05, ** p<.01, *** p<.001

Regarding the second part of the hypothesis, i.e., choosing in accordance with advice when the pie chart is at chance, we did not find the anticipated interaction effect. This would have been an interesting result because going against advice when the non-social cue does not provide any information can be viewed as a clear choice against the information the adviser provides. However, neither the group-by-frame interaction nor the main effect of group were significant, whereas the main effect of frame was (df=(1,150), F=9.645, p=0.002, η^2=0.06, see Supplementary Table 2).

8.2.2.2 Discussion of debriefing questionnaire related effects not reported in the main paper

Additionally, we expected the same group-by-frame interaction effects, where the frame influences low PD participants than high PD participants for the task-dependent debriefing questionnaire measures, namely items 10, 11.2, 11.3, 11.5, 13, and 14. Regarding the debriefing questionnaire measures, none of the anticipated interactions were significant. We also stated anticipating main effects in the analysis plan. However, those main effects are linked to the remaining hypotheses II-VI; please refer to the respective sections in the main text of the paper. For the results of the remaining ANOVAs on debriefing questionnaire items, see Supplementary Table 3.

ANOVA Table for **item 10:**

After the practice round, what expectations did you have with regard to the adviser's behaviour in the upcoming task? Did you expect him/her to give you rather helpful/correct or misleading/incorrect advice?

helpful/correct advice (1) misleading/incorrect advice (6)

0----------0----------0----------0----------0----------0

	SumSq	***Df***	***F***	***p***
Framing Condition	22.046	1	17.384	5.196e-05 ***
Group	14.868	1	11.724	7.997e-04 ***
Group x Frame (IA)	2.945	1	2.323	0.13
Error (within groups)	186.42	147	-	-
Total	225.8	150	-	-

ANOVA Tables for **answer options of item 11:**

The adviser did not always give you helpful/correct advice. What role do you believe the adviser played in this? ?

Please indicate how much you consider the statements below to be true:

	Endorsement of statement 11.2: „The adviser intentionally gave false information, without profiting from it."			
	SumSq	***Df***	***F***	***p***
Framing Condition	0.342	1	4.619	0.033 *
Group	1.251	1	16.866	6.632e-05 ***
Group x Frame (IA)	0.0004	1	0.005	0.943
Error (within groups)	10.899	147	-	-
Total	12.506	150	-	-

	Endorsement of statement 11.3: „The adviser intentionally provided incorrect advice in order to prevent me from scoring points."			
	SumSq	***Df***	***F***	***p***
Framing Condition	2.168	1	27.393	5.629e07 ***
Group	0.872	1	11.014	0.001 **
Group x Frame (IA)	0.003	1	0.032	0.859
Error (within groups)	11.636	147	-	-
Total	14.722	150	-	-

	Endorsement of statement 11.4: „The adviser intentionally provided false information because this was part of his/her instructions/task."			
	SumSq	***Df***	***F***	***p***
Framing Condition	2.963	1	31.651	9.038e-08 ***
Group	0.332	1	3.903	0.050
Group x Frame (IA)	0.010	1	0.114	0.736
Error (within groups)	12.508	147	-	-
Total	15.595	150	-	-

	Endorsement of statement 11.5: „The adviser, in general, did not have full information and made mistakes."			
	SumSq	***Df***	***F***	***p***
Framing Condition	1.358	1	13.643	3.111e-04 ***
Group	0.001	1	0.014	0.907
Group x Frame (IA)	0.228	1	2.291	0.132
Error (within groups)	14.626	147	-	-
Total	16.304	150	-	-

ANOVA Table for **choosing answer option a) of item 13:**
When the adviser provided you with misleading/incorrect advice: What do you believe was the cause of this? *Please assign percentages to the different aspects such that you reach a sum of 100% in total (you can also assign 0%).*
a) The adviser you were assigned
b) The rules of the game / the game's setup

	SumSq	***Df***	***F***	***p***
Frame	0.065	1	0.960	0.329
Group	0.400	1	5.900	0.016 *
Group x Frame (IA)	0.002	1	0.030	0.863
Error (within groups)	9.976	147	-	-
Total	10.447	150	-	-

ANOVA Table for **choosing answer option b) of item 14:**
In your opinion, what factors determined your performance in the task?
Please assign percentages to the different aspects such that you reach a sum of 100% in total (you can also assign 0%).
a) You as the player
b) The adviser which was appointed to you
c) The rules / the structure of the game

	SumSq	***Df***	***F***	***p***
Frame	0.018	1	0.482	0.489
Group	0.023	1	0.590	0.444
Group x Frame (IA)	0.028	1	0.719	0.398
Error (within groups)	5.628	147	-	-
Total	5.693	150	-	-

Supplementary Table 3 | **ANOVA with interaction term on debriefing items.**
N=151, significance levels: * p<.05, ** p<.01, *** p<.001

Furthermore, we analysed participants' agreement with the statement in debriefing questionnaire item 11.4 „The adviser intentionally provided false information because this was part of his/her instructions/task.". Debriefing questionnaire items 11.4 and 11.5 essentially test if the framing-specific instructions influenced the way participants attributed the causes of receiving incorrect advice. Since in the dispositional frame we highlighted the possibility of incorrect advice being caused by the adviser, debriefing questionnaire item 11.4 should generally be endorsed more in the dispositional framing compared to the situational framing condition. The opposite should be the case for debriefing questionnaire item 11.5 ("The adviser, in general, did not have full information and made mistakes.") as it probes the information highlighted in the situational frame, namely the possibility that incorrect

advice was caused by the adviser not having full information. For both debriefing questionnaire items, the main effect of framing was significant in the anticipated direction (Q11.4: df=(1,150), F=31.651, p=9.04e-08, η^2=0.18 and Q11.5: df=(1,150), F=13.643, p=3.11e04, η^2=0.08, see Supplementary Table 3). Additionally, we did not find a significant main effect of group nor a significant interaction. Thus, the framing instructions seemed to influence participants' attributions of incorrect advice after the task similarly across both groups.

8.2.3 Hypothesis II: High PD participants attribute misleading advice validity to the adviser's character rather than to the possibility of the adviser making a mistake due to incomplete information.

We expected participants in the high PD group to attribute misleading advice less to the possibility of the adviser making a mistake due to incomplete information than participants in the low PD group.

8.2.3.1 Discussion of task-behaviour related effects not reported in the main paper

As shown in Figure 1 of the main paper, we designed the input structure such that advice validity was high at the beginning of the task. This phase was followed by a volatile period, after which we presented a similarly structured stable, high advice validity phase. This represents a third (within-subject) factor to our design (phase: low volatility vs. high volatility). We expected that any *a priori* beliefs about potentially malevolent intentions of the adviser would make it more likely that the perceived volatility is attributed to wilful decisions rather than to lack of information. Such beliefs likely occurred in both framing conditions for the high PD group, but only in the dispositional frame for the low PD group. Therefore, we predicted a three-way (group-by-frame-by-phase) interaction, i.e., we anticipated that in the situational frame, high PD scorers would go more frequently against the advice when it is volatile, and that this framing-specific difference between volatile and stable phases is stronger than in low PD scorers. In the dispositional frame, we expected smaller (if any) group differences. However, the three-way ANOVA did not reach significance (see Supplementary table 4).

	Three-Way ANOVA on **taking advice in volatile vs. stable helpful phases**			
	SumSq	***Df***	***F***	***p***
Intercept x Phases (volatile vs. stable)	1.527	1	266.47	7.946e-35 ***
Frames x Phases	0.001	1	0.152	0.697
Group x Phases	0.022	1	3.863	0.051
Frames x Group x Phases	0.002	1	0.290	0.591
Error (Phases)	0.842	147	-	-

	Three-Way ANOVA on **taking advice in stable helpful phase 1 vs. stable helpful phase 2**			
	SumSq	***Df***	***F***	***p***
Intercept x Phases (stable 1 vs. 2)	0.105	1	15.313	0.0001 ***
Frames x Phases	0.006	1	0.808	0.228
Group x Phases	0.002	1	0.364	0.577
Frames x Group x Phases	0.006	1	0.819	0.254
Error (Phases)	1.008	147	-	-

Supplementary Table 4 | **Three-way Repeated Measures ANOVAs with interaction term on task behaviour.**
N=151, significance levels: * p<.05, ** p<.01, *** p<.001

8.2.3.2 Discussion of debriefing questionnaire related effects not reported in the main paper

In the analysis plan, we stated that we expected a significant interaction and main effect of group regarding debriefing questionnaire item 11.5 ("*The adviser, in general, did not have full information and made mistakes.",* percentage assignment from 0% to 100%), where participants were asked to rate how much this statement applies. However, only the main effect of the frame was significant (see Table 3). This makes sense insofar as that one way of inducing the framing was done via reminding participants in the situational frame, that "the adviser does not have full information and can make mistakes." Having a main effect of the frame here thus indicates that all participants trusted the instructions, despite their behaviour showing otherwise in the high PD group. It may well be the case that high PD participants' meta-cognition about their task performance deviated from their behaviour.

8.2.4 Hypothesis III: High PD compared to low PD participants generally attribute negative events to more external-personal causes.

Participants in the high PD group are expected to generally attribute negative events more to external-personal causes as opposed to external-situational causes. We assumed that high PD participants would exhibit higher other-person attribution scores (P-scores) than situation attribution scores (S-scores) across both experimental frames as well as higher P-Scores in general compared to low PD participants.

These scores were based on the internal, personal and situational attributions questionnaire (IPSAQ, see (Kinderman & Bentall, 1996) for more information), which was structured similarly to the debriefing items analysed here.

8.2.4.1 Discussion of debriefing questionnaire related effects not reported in the main paper

We assumed that this attribution style might be expressed by participants' ratings of what/who was responsible for their performance in the task. For participants who were not satisfied with their performance, the following group-by-frame interaction was expected: The low PD group should attribute their unsatisfactory performance more to the situation ('the rules of the game') in the situational frame compared to the dispositional frame, and the high PD participants should attribute poor performance more to the adviser across both frames.

We tested this with debriefing question 14 *("In your opinion, what factors determined your performance in the task?"–* with focus on items b and c). Because the expected interaction is related to a subjective perception of unsatisfactory performance, we only included participants who rated debriefing question 2 (*"How happy are you personally with your performance on this task",* 7-point Likert scale) with 4 or lower (*N*=92).

In order to test these hypotheses related to the above-mentioned P-scores and S-scores, i.e., the odds ratio of the scores, see below:

$$P-score = \frac{mean\ other\ person\ attribution\ (Q14.b)}{1-\ mean\ other\ person\ attribution\ (Q14.b)}$$

$$S-score = \frac{mean\ situation\ attribution\ (Q14.c)}{1\ -\ mean\ situation\ attribution\ (Q14.c)}$$

The way these scores are computed are based on a questionnaire structured similarly to debriefing question 14 (Kinderman & Bentall, 1996). Note however, that computing these scores can lead to numerical problems for those cases in which the divisor equals zero. To avoid these "division by zero" problems, we applied a numerical solution by subtracting an ε-offset from 1 (with ε=10^{-4}). Furthermore, as the data were non-normally distributed, we transformed the numbers logarithmically. Neither the main effects nor the interaction reached significance for these indices (see Supplementary Table 5). The same pattern was observed when analysing all participants without the restriction of conditioning item 14 responses on item 2, or when refraining from using P- and S-scores and only analysing the endorsement of answer option b or c in item 14. This might be due to the fact that most participants attributed failure to themselves rather than the other two.

	ANOVA Table for **P-Scores of Debriefing Item 14**			
	SumSq	***Df***	***F***	***p***
Frame	0.003	1	0.373	0.543
Group	0.003	1	0.400	0.529
Group x Frame (IA)	0.023	1	2.598	0.111
Error (within groups)	0.766	88	-	-
Total	0.796	91	-	-

	ANOVA Table for **S-Scores of Debriefing Item 14**			
	SumSq	***Df***	***F***	***p***
Frame	0.020	1	1.068	0.304
Group	0.005	1	0.282	0.597
Group x Frame (IA)	0.002	1	0.081	0.777
Error (within groups)	1.639	88	-	-
Total	1.665	91	-	-

Supplementary Table 5 | **Two-way ANOVA with interaction term on debriefing item 14.**
N=92 participants who expressed dissatisfaction with their performance. Answers of this subset of participants on debriefing item 14: *" In your opinion, what factors determined your performance in the task?"*, with the P-Score representing agreement with answer option *"the adviser which was appointed to you"* normalised by the other answer options and the S-Score representing agreement with answer option *"the rules / the structure of the game"* normalised by the other answer options

Since the planned ANOVA is not the most appropriate test for debriefing questionnaire item 14 and in light of the data being non-normally distributed, we ran a Wilcoxon rank sum test on the answers to answer option b (*"the adviser which was appointed to you"*) and also on the answers to answer option c (*"the rules / the structure of the game"*). However, this test did not show any significant group differences either (P-Score: z=-0.24, p=0.60; S-Score: z=--0.41, p=0.34).

8.2.5 Hypothesis IV: High PD participants attribute differences in advice validity to the adviser being malevolent.

We expected high PD participants to attribute changing advice validity to the adviser's character, i.e., a general tendency of the adviser to be malevolent.

8.2.5.1 Discussion of debriefing questionnaire related effects not reported in the main paper

High PD participants were expected to score higher than low PD participants on debriefing questionnaire item 11.2 (*"The adviser intentionally gave false information, without profiting from it."*, percentage assignment from 0% to 100%) in the situational frame, but not in the dispositional frame (see Hypothesis I.), implying an interaction governed by the experimental frame.

While this interaction did not reach significance, the main effect of group did, which is reported in the main text of the paper. The anticipated main effect of group in item 13 (rating the adviser as the cause

of incorrect advice being more frequent in the high PD group than the low PD group) was significant and reported in the main text (see also Supplementary Table 3).

Additionally, we asked participants to rate how likable and how warm they perceived the adviser to be in the beginning (Debriefing questionnaire item 15) of the task and how the assessed the adviser after the task (Debriefing questionnaire item 16). Under the assumption that individuals in the high PD group would more readily attribute false advice to the adviser's character, experiencing false advice during the task should impact their likability ratings. Consistent with this prediction, we found that high PD participants' ratings between item 15 and 16 differed significantly more than low PD participants' ratings (two-tailed *t*-test; df=149, t=2.47, p=0.01, *Cohen's d*=0.4).

8.2.6 Hypothesis V: High PD participants should expect receiving misleading advice.

High PD participants should be less surprised by misleading advice, as to them misleading advice is more likely than helpful advice. High PD participants should therefore expect misleading advice before playing the game. As a consequence, they should report expecting misleading advice in the initial behavioural probe more so than participants in the low PD group. This effect is also expected to be moderated by the experimental frames.

8.2.6.1 Discussion of task-behaviour related effects not reported in the main paper

We found no significant main effect of group or interaction effect on the initial behavioural probes when running the planned ANOVA (two-way ANOVA; main effect of group: df=(1,150), F=0.29, p=0.58; interaction (group × frame): df=(1,150), F=0.07, p=0.79). The main effect of frame, however, was significant (df=(1,150), F=43.15 p=0.003). Because the answer options on the task-probe are on a nominal scale and an ANOVA is not the appropriate test, we decided to test this main effect with a Kruskal-Wallis test for non-parametric data. Indeed, the main effect of the frame was significant (df=149, χ^2= 7.56, p=0.006, Φ =0.22), indicating that the initial assessment of the adviser's fidelity was significantly influenced by the framing, with a more negative assessment of the adviser in the dispositional compared to the situational frame, irrespective of participant group.

Additionally, we expected a three-way group × frame × phase interaction for taking helpful advice in the first and the last stable helpful phase (right at the beginning of the task and at the very end of the task, after the volatile phase). High PD participants were expected to show a lack of adaptability to the frame and therefore take advice less in stable helpful phases, especially in the first stable helpful phase compared to the last helpful phase. This interaction, however, did not reach significance (see Supplementary Table 4 and Figure 3c in the main text). The significant main effect of phase showed that all participants across framing conditions took advice less into account in the stable helpful phase at the end of the task (following the volatile phase) compared to the same phase occurring at the beginning of the task (df=(1,147), F=103.42, p<0.0001, $\eta^2_{partial}$=0.09, including Greenhouse-Geisser nonsphericity correction).

8.2.7 Hypothesis VI: High PD participants view incorrect advice as directed towards them.

High PD participants should not only view incorrect advice as expected behaviour exhibited by others, but we also expected them to believe that such behaviour was directed towards themselves.

8.2.7.1 Discussion of debriefing questionnaire related effects not reported in the main paper

We tested this hypothesis with debriefing questionnaire item 11.3 *("The adviser intentionally provided incorrect advice in order to prevent me from scoring points.",* percentage assignment from 0% to 100%), with higher ratings for high PD participants versus low PD participants in the situational frame but not in the dispositional frame The interaction was not significant (see Supplementary Table 3), whereas the main effect of group was. This result was reported in the main text of the paper, and labelled as a post-hoc analysis. The main effect of frame was also significant.